WHAT IS CLAIMED IS:

1. A method in a natural language processing system of segmenting a textual input string including a plurality of characters arranged in character groups separated by white spaces, the method comprising:

receiving the input string;

segmenting the input string into one or more proposed tokens;

validating the proposed tokens by submitting the proposed tokens to a linguistic knowledge component to determine whether the proposed tokens represent linguistically meaningful units; and

if not, re-segmenting the input string into one or more different proposed tokens.

2. The method of claim 1 wherein segmenting comprises:

accessing segmentation criteria arranged in a predetermined hierarchy of segmentation criteria, and segmenting based on the segmentation criteria in an order based on the hierarchy.

3. The method of claim 2 wherein segmenting according to the hierarchy of segmentation criteria comprises:

accessing language-specific data containing a portion of the segmentation criteria.

4. The method of claim 3 wherein segmenting comprises:

accessing a precedence hierarchy of punctuation in the language-specific data, the precedence hierarchy being arranged based on binding properties of the punctuation in the precedence hierarchy, and segmenting the input string based on the punctuation in an order based on the precedence hierarchy.

5. The method of claim 2 and further comprising:

repeating the steps of validating and re-segmenting until all characters in the input string have been validated or until the predetermined hierarchy of segmentation criteria has been exhausted.

6. The method of claim 1 wherein the linguistic knowledge component includes a lexicon and wherein validating comprises:

accessing the lexicon to determine whether it contains the proposed tokens.

7. The method of claim 6 wherein the linguistic knowledge component includes a morphological analyzer and wherein validating comprises:

invoking the morphological analyzer to convert a form of the proposed tokens to a morphologically different form; and

accessing the lexicon to determine whether it contains the morphologically different form of the token.

8. A segmenter segmenting a textual input string, containing characters, into linguistically meaningful units, the segmenter comprising:

a data store storing language-specific data indicative of a precedence hierarchy of punctuation arranged based on binding properties of the punctuation;

a linguistic knowledge component configured to validate a token as a linguistically meaningful unit; and

an engine coupled to the data store and the linguistic knowledge component and configured to receive the input string, access the language-specific data in the data store, segment the input string into one or more proposed tokens, submit the one or more proposed tokens to the linguistic knowledge component for validation, and if the linguistic knowledge component is unable to validate the one or more proposed tokens, re-segment the input string into one or more different proposed tokens.

9. The segmenter of claim 8 wherein the engine is further configured to repeatedly re-segment the input stream into one or more different proposed tokens based on a predetermined hierarchy of segmentation criteria, and resubmit the one or more different proposed tokens to the linguistic knowledge component until the segmentation criteria are exhausted or all the characters in the input string are validated.

10. A method of segmenting a textual input string including characters separated by spaces, comprising:

receiving the textual input string;

proposing a first segmentation of at least a portion of the input string;

attempting to validate the first segmentation by submitting the first segmentation to a linguistic knowledge component; and

if the first segmentation is not validated, proposing a subsequent segmentation and submitting the subsequent segmentation to the linguistic knowledge component for validation.

11. The method of claim 10 and further comprising:

repeating the steps of proposing a subsequent segmentation and submitting the subsequent segmentation to the linguistic knowledge component until the portion of the input

string is validated or the portion of the input string has been segmented according to a predetermined number of segmentation criteria.

12. The method of claim 11 wherein proposing a first segmentation comprises:

segmenting the input string at the spaces to obtain a plurality of tokens.

13. The method of claim 12 wherein proposing a subsequent segmentation comprises:

determining whether invalid tokens contain any of a predetermined plurality of multi-character punctuation strings or emoticons; and

if so, segmenting the tokens into subtokens based on the multi-character punctuation strings or emoticons .

14. The method of claim 13 wherein proposing a subsequent segmentation comprises:

determining whether invalid tokens contain punctuation marks; and

if so, segmenting the tokens into subtokens according to a predetermined precedence hierarchy of punctuation.

15. The method of claim 14 wherein proposing a subsequent segmentation comprises:

determining whether invalid tokens contain both alpha and numeric characters; and

if so, segmenting the tokens into subtokens at boundaries between the alpha and numeric characters in the tokens.

16. The method of claim 15 wherein proposing a subsequent segmentation comprises:

reassembling previously segmented subtokens.

17. The method of claim 11 wherein proposing a first segmentation comprises:

identifying a token as a group of characters flanked by spaces or either end of the input string.

18. The method of claim 17 wherein proposing a subsequent segmentation comprises:

determining whether the token contains either all alpha characters or all numeric characters; and

if so, indicating that the token cannot be further segmented and will be treated as an unrecognized word.

19. The method of claim 18 wherein proposing a subsequent segmentation comprises:

determining whether the token includes final punctuation; and

if so, segmenting the token into a subtoken by splitting off the final punctuation.

20. The method of claim 19 wherein proposing a subsequent segmentation comprises:

determining whether the token includes both alpha and numeric characters; and

if so, segmenting the token into subtokens at a boundary between the alpha and numeric characters.

21. The method of claim 20 wherein proposing a subsequent segmentation comprises:

determining whether the token includes one or more of a predetermined set of multi-punctuation characters or emoticons; and

if so, segmenting the token into subtokens based on the multi-punctuation characters or emoticons included in the token.

22. The method of claim 21 wherein proposing a subsequent segmentation comprises:

determining whether the token includes one or more edge punctuation marks; and

if so, segmenting the token into subtokens by splitting off the one or more edge punctuation marks according to a

predetermined edge punctuation precedence hierarchy.

23. The method of claim 23 wherein proposing a subsequent segmentation comprises:

determining whether the token includes one or more internal punctuation marks, internal to the tokens; and

if so, segmenting the token into subtokens based on the one or more internal punctuation marks according to a predetermined internal punctuation precedence hierarchy.